IBM Data Science Professional Certification by Coursera

Applied Data Science Capstone

Opening a New Multiplex Business in Hyderabad, India

Phalgun Reddy Bobbili

August 2020

Introduction

The Indian film industry is the produces the greatest number of movies every year than any other film industry. Every year, around 2000 movies are released across many languages in the country. In 2015, the total box office gross was over $2.1 billion, third highest in the world. The people of India love going to the cinemas to enjoy the experience of a big screen viewing. Even when the OTT platforms are rising and gaining users and traction in the country, the box office collections have not decreased. This is because the public still prefer to go to the cinema theatres for the that big screen experience. This is even more applicable to the southern part of the country, where the love for cinemas even more special. Hyderabad is one of the most developed metropolitan cities in the country and located in southern India.

There are basically two kinds of cinema theatres. The single screen theatres and the multiplexes which offer multiple screens. The pubic these prefer going to multiplexes over the single screen theatres. The multiplexes come with a various shopping opportunities and food courts. Also, the multiplexes offer multiple movies at many different times of the day, allowing the people to choose their preferred movie at their preferred time. These factors enable the people to plan a whole evening at one location, attracting a lot of customers. On the other hand, multiplexes require huge investments to be established. Hence, a lot of planning should go into it. One of the most important aspects of a multiplex that drive its success is the location of the multiplex. Factors like the number of competitors in the area, the demographics of the area, etc. influence the success of the business heavily.

Business Problem

The objective of this capstone project is to analyze the city of Hyderabad and try to select the most suitable locations to open a new multiplex business in the city. This project aims to achieve this by using data science methodologies and machine learning techniques like clustering.

Data

To successfully finish this project, we will need the following data:

• We will be analyzing the various locations of the city by neighborhoods. Hence, we will need a list of neighborhoods in Hyderabad.

• We will use the coordinates of the neighborhoods to plot Folium maps of the city. We will need the latitudes and the longitudes of the neighborhoods for this.

• We will retrieve venues data from Foursquare API and extract the multiplexes out of the venues data. Venue data, particularly data related to Multiplexes. This data will be used to cluster the neighborhoods.

Methodology

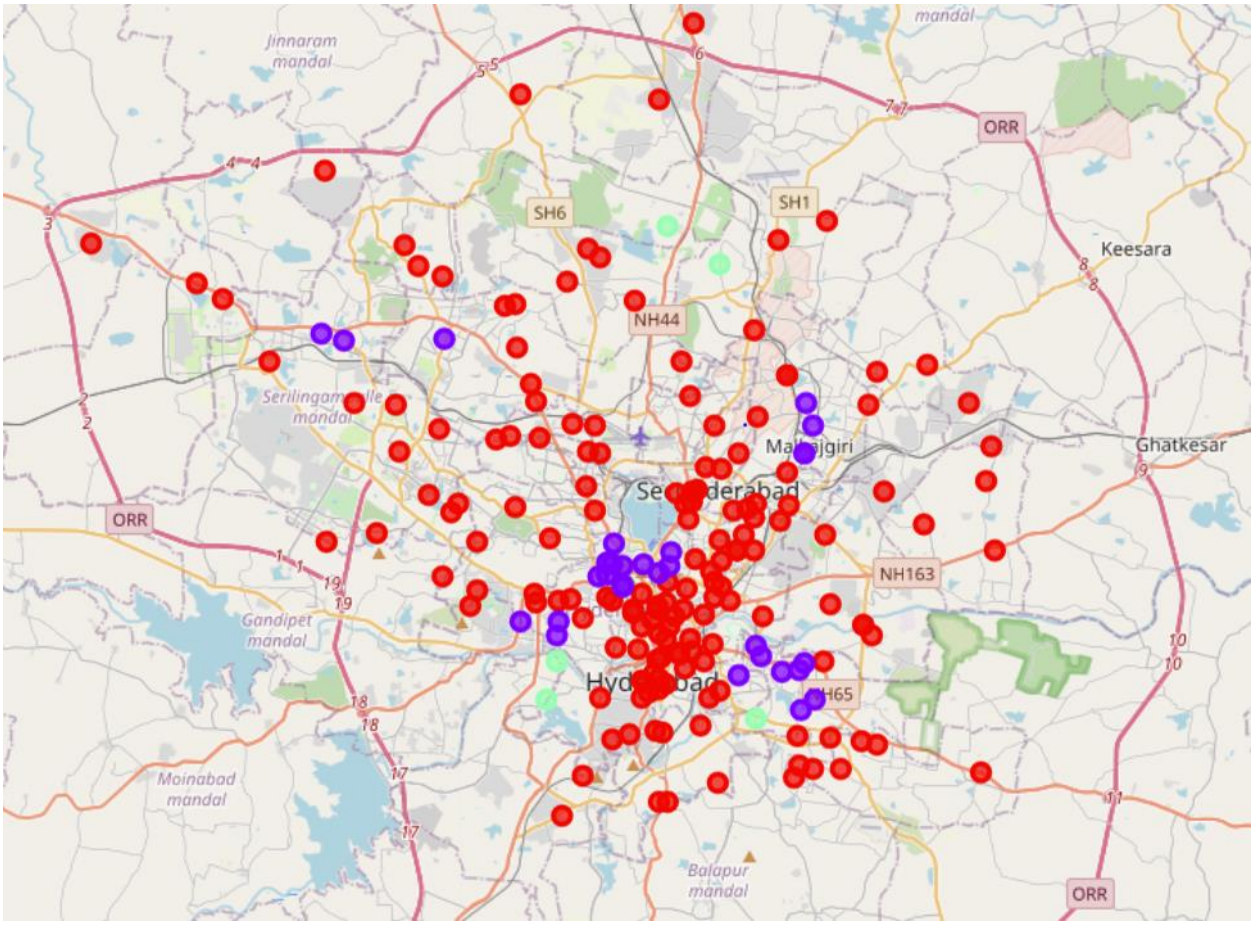
We begin with a list of neighborhoods in the city of Hyderabad. We will use the webpage https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India to scrape using python requests and BeautifulSoup packages to extract the list of neighborhoods. We then need the geographical coordinates of these neighborhoods to be able to work with the Foursquare API or plot the folium map. We get this information using the geocoder package in python which converts address into geographical coordinates. We turn the list of neighborhoods with their corresponding coordinates into a dataframe and plot the coordinates in a folium map to visualize. Before we proceed any further, we need to create a Foursquare API account and know the client-Id and client-secret. We will use the Foursquare API to retrieve 100 venues inside a specified radius of 2000. We do this by using a loop to pass the coordinates to the Foursquare API. The venues data is retrieved in JSON format, from which we extract the venue name, category, latitude and longitude. With this information we can analyze the number of venues we got per neighborhood or how many types of categories of venues are there in the data. We will then analyze the neighborhoods by grouping and indexing rows by the neighborhood name and provide the mean of the frequency of occurrence of each venue category by using one-hot encoding. By doing this we have also started the process of preparing the data for clustering. Since our focus category is "Multiplex", we will filter the data for that category. We will now perform k-means clustering. K-means is one of the simplest and popular unsupervised learning algorithms. We distinguish the neighborhoods into three clusters. We will talk about the clusters more in the Results section.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Multiplex":

- Cluster-0: Neighborhoods with moderate number of Multiplexes
- Cluster-1: Neighborhoods with low to no existence of Multiplexes
- Cluster-2: Neighborhoods with high concentration of Multiplexes

The results of the clustering are visualized in the map below with cluster-0 in red color, cluster-1 in purple color, and cluster-2 in mint green.



Discussion

As we can see in the map in the Results section, a big majority of the neighborhoods are of cluster-0 (high concentration of multiplexes). There are a significant number of neighborhoods of cluster-2 (moderate concentration of multiplexes), while there are very few neighborhoods of cluster-1 (few to no multiplexes). We can observe a heavy concentration of multiplexes in the central parts of the city. If we were to judge from just this analysis, we can say that the neighborhoods of cluster-1 may be good choices for establishing a new multiplex in as there is little to no competition in those neighborhoods and the neighborhoods of cluster-0 seem to have saturated competition and choosing those neighborhoods to establish a multiplex in will probably prove to be a wrong decision.

Limitations and Suggestions for Future Research

In this project, we only considered one factor i.e., the frequency of occurrence of multiplexes in the neighborhoods. But realistically thinking, a business decision as big as establishing a new multiplex business cannot be made based just on this analysis. We need to consider many more factors like the population density of the neighborhood, the distribution of rich and poor in that population, the real estate value, etc. But I believe I would be limited by my skill in data science to embark on such an elaborate and complex analysis. Also, the data required for the analysis may be difficult to find in the neighborhood level. We can dive deeper into this analysis in the future after learning more about data science and may be, with an upgraded Foursquare API account with lesser limitations.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters statistically, and lastly providing recommendations to the relevant stakeholders regarding the best locations to open a new Multiplex business. To answer the business question that was raised in the Business Problem section, the answer proposed by this project is: the neighborhoods in cluster 1 are the most preferred locations to open a new multiplex. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new multiplex.

References

Source neighborhoods data:

https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India

Foursquare developer documentation:

https://developer.foursquare.com/docs/